

Exhibit 11

Epidemiology of Commonly Used Statistical Terms and Analysis of Clinical Studies

Wendy R. Brewster, MD, PhD

OUTLINE

Epidemiology
Evidence-Based Medicine
Measures in Epidemiology
Analysis of Clinical Trials
Types of Clinical Trials
Evaluation of Clinical Trials

Placebo Treatment Groups
Controls Used in Clinical Trials
Studies of Therapy
Blinding
When to Stop a Clinical Trial

EXHIBIT 7
WIT: CLARKE-
PEARSON
DATE: 2-4-19
Sophie Brock, RMR, CRR

KEY POINTS

1. Epidemiology is the study of distribution of disease and factors that determine disease occurrence in populations.
2. As much as possible, medical decisions should be based on quality evidence. The best evidence is a properly designed randomized controlled trial. Evidence from

nonrandomized but well-designed control trials is of lesser quality. Next in reliability is well-designed cohort or case-control studies, which have been repeated by several investigators. Opinions of respected authorities and extensive clinical experience are least reliable.

EPIDEMIOLOGY

Epidemiology is the study of distribution of disease and the factors that determine disease occurrence in populations. The focus is on groups rather than the individual. Persons within a population do not have equal risk for disease occurrence, and the risk of a disease is a function of personal characteristics and environmental exposures. Patterns of disease occurrence within specific populations can be evaluated to determine why certain groups develop illness when others do not. The impact of epidemiology on gynecologic oncology is evidenced by the significance of studies such as the association with infection of oncogenic human papillomavirus and cervical cancer, obesity and the risk of endometrial cancer, and the risk factors for gestational trophoblastic neoplasia. Epidemiologic studies are unique in their focus on human populations and their reliance on nonexperimental observations. Epidemiologic methods are used in searching for causes of disease, disease surveillance, determining the cause of disease, diagnostic testing, searching for prognostic factors, and testing new treatments.

Because the quality of epidemiologic evidence varies greatly among studies, the scientific community endorses the principles of Sir Austin Bradford Hill, an eminent British statistician, when attempting to identify causal associations. A cause of a specific disease is an antecedent event or characteristic that is necessary for the occurrence of the disease (Box 22.1).

EVIDENCE-BASED MEDICINE

As much as possible, medical decisions should be based on quality evidence. The best evidence is a properly designed randomized controlled trial. Evidence from nonrandomized but well-designed control trials is of lesser quality. Next in reliability is a well-designed cohort or case-control studies, which have been repeated by several investigators. Opinions of respected authorities and extensive clinical experience are least reliable.

Physicians are currently encouraged to practice evidence-based medicine. This means that clinical trial evidence must pass statistically valid tests for conclusions to have meaning. Good science depends on accurate (ie, statistically significant and meaningful) data from clinical trials. The best trials are usually experimental, powered, randomized, and blinded. Patients randomly assigned to a treatment group or a control group must have an equal probability of being assigned to either group. This prevents selection bias (eg, putting healthier or better prognosis patients in one group and those with a poor prognosis or high likelihood of disease risk in another group). Blinding prevents patients, investigators, or statisticians from knowing who is in the control group and experimental group; thus, biased actions are avoided.

Whereas retrospective and observational studies are descriptive and do not involve either an intervention or a manipulation, an experimental study does. A prospective trial poses the

BOX 22.1 Strength of Association

1. **Temporality:** Exposure must precede the onset of the disease.
2. **Dose-response:** Risk increases as exposure increases.
3. **Replication:** The association is observed repeatedly.
4. **Coherence:** The association is consistent with other scientific knowledge and does not require that established facts be ignored.
5. **Exclusion of the role of chance:** Appropriate statistical tests demonstrate that the observed association is extremely unlikely to have arisen by chance.

Modified from Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295.

TABLE 22.1 Mathematical Definitions of Statistical Terms

| Terminology | Mathematical Definition |
|---------------------------|---|
| Prevalence rate | Number of persons with disease/Total number in the group |
| Incidence rate | Number of new cases/Total number at risk per unit of time |
| Kappa | $(P_{obs} - P_{chance}) / (1 - P_{chance})$ |
| Sensitivity | True positive/(True positive + False negative) |
| Specificity | True negative/(True negative + False positive) |
| Predictive value positive | True positive/(True positive + False positive) |
| Predictive value negative | True negative/(True negative + False negative) |

question before the data are collected, thus allowing better control of confounding variables, unlike a retrospective study, which poses the question after the data are collected.

MEASURES IN EPIDEMIOLOGY

To describe and compare groups in a meaningful manner, it is important to find and enumerate appropriate denominators and statistical terms (Table 22.1).

- **Incidence rate:** Measures the new cases of a specific disease that develop during a defined period of time and the approximation of the risk for developing the disease. The incidence rate focuses on events. Incidence measures the probability of developing a disease.
- **Kappa coefficient:** Kappa indicates how much observers agree beyond the level of agreement that could be expected by chance. Kappa is estimated as $(P_{obs} - P_{chance}) / (1 - P_{chance})$. Thus, the kappa coefficient is the observed agreement, corrected for chance as a fraction of the maximum obtainable agreement, also corrected for chance. Landis and Koch have suggested useful categorizations. Kappa = 0.00 should be taken as representing "poor" agreement, 0.00 ± 0.20 as "slight" agreement, 0.21 ± 0.40 as "fair" agreement, 0.41 ± 0.60 as "moderate" agreement, 0.61 ± 0.80 as "substantial" agreement, and 0.81 ± 0.99 as "almost perfect" agreement. A kappa coefficient of 1 represents perfect agreement.
- **Mean:** The average of a sample of observations.
- **Median:** The middle value when the values are arranged in order from the smallest to the largest.

- **Meta-analysis:** The statistical process of pooling the results from separate studies concerned with the same treatment or issue is frequently used in the context of medical statistics and provides the quantitative backbone of the evidence-based medicine program. A large number of meta-analyses are undertaken with the broad aim of combining divergent outcomes into a single estimate of treatment effect. For example, the Cochrane Collaboration endeavors to collate and synthesize high-quality evidence on the effects of important health care interventions for a worldwide, multi-disciplinary audience and publishes them in the *Cochrane Database of Systematic Reviews*. Meta-analyses increase the statistical power by increasing the sample size, resolve uncertainty when reports do not agree, and improve the estimates of effect size. The bias of publication only of positive results is a concern for those using results of meta-analyses because, if statistically significant or "positive" results are more likely to be published, a meta-analysis based on the resulting literature will be biased. The quality of the studies included is important to the quality of the final result.
- **Pearson's correlation r :** The degree to which two variables are related is called correlation. Pearson's correlation is represented by the value r and varies between -1 and $+1$. It is usually presented as a scatter point graph. A value of -1 suggests a perfect negative linear relationship, a value of 0 reflects no linear relationship, and a value of 1 reflects a perfect linear relationship. Values of -1 , 0 , and $+1$ are rare.
- **Person time:** The sum of the observation period of risk for the persons in a group being studied.
- **Predictive value positive:** The proportion of positive test results that is truly positive (ie, the probability that someone classified as exposed is truly exposed). This value only refers to positive tests.
- **Predictive value negative:** The proportion of negative test results that is truly negative. The predictive value of a negative test result refers to the proportion of patients with a negative test result who are free of disease. These values, unlike sensitivity and specificity, indicate the reliability of the test in the determination of presence or absence of disease.
- **Prevalence rate:** The amount of disease in a population. Prevalence measures the proportion of diseased individuals at a particular time and represents a snapshot of the disease. Other commonly used terms are *prevalence proportion* and *point prevalence*. It is a measure of status and includes individuals with newly diagnosed disease and those surviving with disease. The numerator is the number of affected individuals in a specific time period. The denominator is the total number of persons in the group. Prevalence rates range between 0 and 1 .
- **Quality-adjusted life year (QALY):** The QALY was developed as an attempt to combine the value length of life and quality of life into a single index number. One year of perfect health is given a value of 1 . Death is given a value less than 0 . A year of less than perfect health will have a value less than 1 . States of health considered worse than death can be argued to have a negative value. The QALY value is

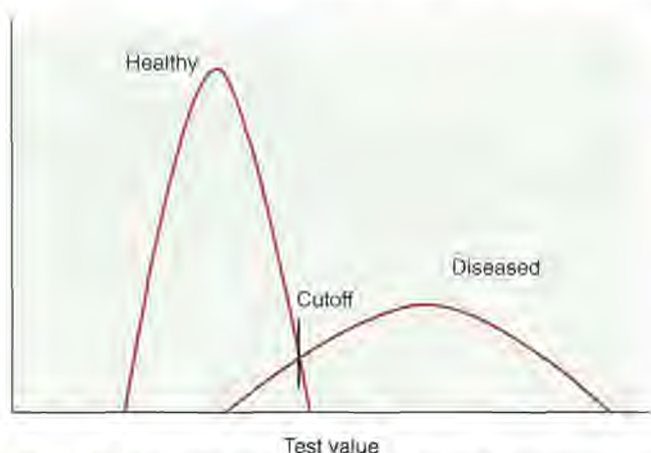


FIGURE 22.1 Effects of shifting cutoff point on sensitivity and specificity.

determined by multiplying the utility value associated with that state of health by the years lived in that state. QALY is a metric used to compare the benefit of health care interventions. Combination of QALYs with the cost of an intervention (Cost/QALY) can provide an economic framework for comparisons of therapies. QALYs have several limitations and should not be used alone in decision making.

- **Sensitivity:** The proportion of truly diseased persons who are classified as diseased by the test. The sensitivity of a test is therefore the probability of a test being positive when the disease is present. The sensitivity of a test may also be called the true-positive rate. In Fig. 22.1, it is evident that the cutoff point of a test can affect the sensitivity. If the cutoff point is moved to the left, more diseased persons will be identified. At the same time, more healthy persons will be erroneously classified as sick. However, as the cutoff value for normal is moved to the right, the test will become less sensitive because fewer diseased persons will be classified as such.
- **Specificity:** The proportion of a population of disease-free individuals who are classified as undiseased by a test. In contrast to the sensitivity of a test, the specificity of a test is the probability that a test result will be negative when the disease is absent. The cutoff point of a test for normality influences the specificity. As the value of normality or cutoff moves to the left, the test becomes less specific because fewer healthy individuals are recognized as such. In contrast, moving the cutoff values to the right increases the specificity (see Fig. 22.1). In the best scenario, a test would be able to discriminate between diseased and healthy individuals without any overlap. More often, the scenario is as presented in Fig. 22.1, in which there is significant overlap and whatever the cutoff value healthy persons may be classified as diseased and sick persons classified as healthy. When we set the cutoff point for a test, we must be attentive to the purpose of the test. If the disease is treatable and missing the disease has serious ramifications, then we must favor sensitivity over specificity. Alternatively, if it is more important to correctly identify healthy individuals, then specificity is prioritized. Published reports of the performance of tests usually just

provide sensitivity and specificity results. Variations of these measures occur under many conditions and will also produce variations in predictive values.

- **Standard deviation (SD):** A measure of the variability within each group. If there is a normal (bell-shaped curve) distribution, approximately 95% of the values are within 2 SDs on both sides of the average.
- **Tests of heterogeneity:** Before performing a meta-analysis, it is customary to assess evidence of variation in the underlying effects. This variation, termed "heterogeneity," arises because of differences across studies in populations, exposures or interventions, outcomes, design, or conduct. A forest plot is useful for visual assessment of consistency of results across studies. A statistic that measures the consistency of findings as the proportion of total variation in point estimates attributable to heterogeneity is now widely used.

ANALYSIS OF CLINICAL TRIALS

- **Null hypothesis:** This hypothesis, symbolized by H_0 , is a statement claiming that there is no difference between the experimental and population means. The alternative hypothesis (H_1) is the opposite of the null hypothesis. Often in research, we need to be able to test for both the positive and adverse outcomes; therefore, a two-tailed hypothesis is chosen even though the expectation of the experiment is in a particular direction.
- **Significance level:** A level of significance termed the alpha value is determined before the study has begun. The alpha value is the likelihood that a difference as large or larger that occurred between the study groups could be determined by chance alone. The alpha level is established by those designing the study and becomes the level of statistical significance. The most typical alpha level is 0.05.
- **One-tail test:** A test to determine a difference in only one direction (eg, to determine if drug A is better than drug B).
- **Two-tail test:** A test to determine any difference between the variable (eg, if either drug A or drug B is superior to the other). It is usually considered that in a two-tailed test, more trust can be placed in the statistically significant results than with a one-tailed test. When in doubt, the two-tailed test is preferred.
- **Confidence interval (CI):** The range of values that is believed to contain the true value within a specific level of certainty.
- **Alpha error:** The rejection of the null hypothesis when it is, in fact, correct; also called a type I error.
- **Beta error:** Failure to reject the null hypothesis when it is, in fact, incorrect; also called a type II error.
- **Power:** The probability that a study will be able to correctly detect a true effect of a specific magnitude. The statistical power refers to the probability of finding a difference when one truly exists or how well the null hypothesis will be rejected. The power is usually specified beforehand in prospective studies. The values of 0.8 (80%) or 0.9 (90%) are typical. The higher the value, the less chance there is of a type II error. A 0.9 value means that a type II error would be avoided 90% of the time.

- **Risk:** The proportion of unaffected individuals who, on average, will contract the disease of interest over a specified period of time. Results of a trial are often expressed as absolute or relative risk reductions. The absolute difference is the actual difference between the units of the difference. In relative risk, the differences are the percentage change. Relative risk reductions often sound much more dramatic than do the absolute values. One must consider the prevalence of a disease when evaluating risk reductions. When there is a low prevalence of a disease process, small risk reductions become unimpressive and must be evaluated in terms of the benefits of a particular mode of therapy.
- **Incremental cost-effectiveness ratio:** The additional cost divided by the incremental benefit compared with an alternate strategy. A strategy was strongly dominated if it was more costly and less effective than another or cost effective if it had an incremental cost-effectiveness ratio of \$50,000 to 100,000 per year of life gained relative to an alternate strategy.
- **Odds ratio (OR):** The ratio of the odds that an event will occur in one group compared with the odds that the event will occur in the other group. In an osteoporosis study, if 14 of 22 people who are thin, have fractures, the odds of having a fracture are 14 in 22 or 0.64. If 5 of the 33 nonthin people fracture bone, the odds are 5 in 33 or 0.15. The OR is 0.64 divided by 0.15 or 4.2, meaning that thin people are 4.2 times more likely to receive fractures. An OR of 1 means that both groups have a similar likelihood of having an event.
- **Overall survival (OS):** The interval from the completion of treatment to censoring or death from any cause.
- **Progression-free survival (PFS):** The interval from the date of randomization to the documentation of progression of the illness or death from any cause.
- **Actuarial (life table) survival:** This technique uses grouped information to estimate the survival curve. The data are grouped into fixed time periods (eg, months, years) that include the maximum follow-up. The survival curve is estimated as a continuous curve and gives an estimate of the proportions of a group of patients who will be alive at different times after the initial observation. The group includes patients with incomplete follow-up.
- **Chi square (χ^2):** The primary statistical test used for studying the relationship between variables. This is a test used to compare proportions of categorical variables.
- **Cox proportional hazard regression analysis:** Cox regression analysis is a technique for assessing the association between variables and survival rate. The measure of risk provided for each variable is the risk ratio (RR). An RR of 1 means that the risk is the same for each participant. An RR greater than 1 indicates increased risk; a ratio less than 1 indicates less risk. A ratio of 5.4 means that the patients with a variable are 5.4 times more likely to have the outcome being studied. CIs can also be provided with RRs. This type of analysis is usually presented in a table.
- **Efficacy:** The possibility that an intervention will result in a change (eg, in vaccine trials).
- **Imputation:** In many analytical scenarios, a case is discarded if it is missing data relevant to the analysis. This may introduce a bias, decrease the power of the study, or alter the representativeness of the results. Imputation is the process of substituting missing data. There are numerous techniques to impute data.
- **Multivariate analysis:** A technique of analysis of data that factors many variables. A mathematical model is constructed that simultaneously determines the effect of one variable while evaluating the effect of other factors that may have an influence on the variable being tested. The two most common algorithms developed to accomplish this task are the step-up and step-down procedures. Variables are added to an initial small set or deleted from an initial large set while testing repeatedly to see which new factor makes a statistical contribution to the overall model.
- **Propensity score:** In studies in which randomization is not possible, investigators use this score to adjust for the bias due to confounding variables inherent in treatment selection for groups being compared (eg, in observational trials when the treatment of the participants is not random). Assignment to intervention does not occur by chance but depends on patient characteristics that can influence the effect of the intervention on the outcome. The propensity score is the probability that a patient would receive the treatment of interest based on the characteristics of the patient, clinician, and clinical environment. Several different techniques can be used to balance the groups being compared. The most common strategy is propensity score matching. Propensity score methods are usually better than matching on specific characteristics or stratification because it adjusts for confounding better than other strategies. This is a valuable approach when completion of a randomized trial is not feasible.
- **Receiver operator characteristics (ROCs):** These curves are the best way to demonstrate the relationships between sensitivity and specificity. The curves plot sensitivity (true-positive rate) against the false-positive rate ($1 - \text{Specificity}$) (Fig. 22.2). The closer the curve is to the upper lefthand corner, the more accurate it is because the true-positive rate is closer to 1 and the false-positive rate is closer to 0. Along any particular ROC curve one can observe the impact of compromising the true-positive and false-positive rates. As the requirement that a test has a high true-positive rate increases, the false-positive rate will also increase. The closer the curve is to the 45-degree diagonal of the ROC area under the curve, the less accurate the test (see Fig. 22.2).
- **Sensitivity analysis:** Modeling is a tool that permits the incorporation of data from different sources to predict outcomes that may be likely. This tool allows for various interventions and numerous scenarios. Interpretation of the findings depends on the confidence the operator has in various factors of the model. Sensitivity analysis involves examination of the variation in the outcome as the variables change.
- **One-way sensitivity analysis:** This is the simplest form of sensitivity analysis in which the effectiveness of only one

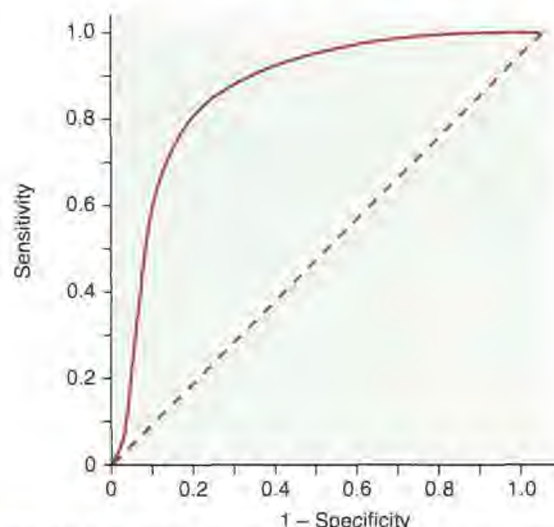


FIGURE 22.2 A hypothetical example of a receiver operating characteristic curve. The solid line represents the performance of the diagnostic test of interest; the dashed diagonal line serves as a reference of a test with no diagnostic value. (From Greenberg RS, Daniels SR, Flanders WD, et al: *Medical Epidemiology*, 3rd ed. Lange Medical Books/McGraw-Hill, 2001 Reproduced by permission of The McGraw-Hill Companies.)

variable in the model is analyzed to assess the range of impact of that variable on the outcome.

- **Univariate analysis:** Analyses may be univariate or multivariate because they examine one or more variables at a time, respectively.

TYPES OF CLINICAL TRIALS

Clinical trials are experiments in which the investigator intervenes rather than observes and is the best test of a cause-and-effect relationship. A properly conducted experiment requires that when the intervention is applied to one group, there is a control group or some other suitable standard by which participants of the clinical experiment or their guardians must give informed consent. The gold standard of clinical trials is the randomized experiment. With randomization, each participant has an equal chance of being in either of the arms of the trials. Randomization is important because it equalizes baseline characteristics of the participants so that the comparison of the treatments is fair. If randomization is not feasible, possible nonrandom standards of comparison must include patients similar to the treated group. Randomization is the current norm for demonstrating efficacy and safety of investigational methods. The advantages of randomization are numerous:

- Decreases investigator's bias in assigning patients to treatment groups
- Permits certain statistical methods to be used in the resulting data
- Allows for blinding of the patient and investigator
- Is the current norm for demonstrating efficacy and safety of investigational medicines

BOX 22.2 Types of Clinical Trials

Single-patient clinical trials
 Multicenter trials
 National clinical trials
 Continuation trials
 Compassionate plea trials
 Pharmacoeconomic trials
 Trials to evaluate medical devices
 Pharmacokinetic trials

Unacceptable methods of randomization include the following:

- Alternate assignments
- Alternate day assignments
- Birthday assignments
- Coin tosses
- Initials of a patient

Single-patient clinical trials are indicated only in specific situations. They are generally used to evaluate rare diseases when other types of trials are inappropriate or when only a small percentage of patients respond to a specific treatment. Single-patient clinical trials are useful to determine the response of a particular patient is a result of placebo or if an adverse reaction is related to a specific medication. The disease should be chronic, and the disease severity must be stable during the clinical trial duration. It should be expected that the effect of the treatment should be measurable in a short time, and the effect should be rapidly reversible after the treatment has stopped. The investigator and patient should be blinded, and the patient's condition should return preexisting baselines between treatment legs.

Multicenter trials are advantageous because they offer more rapid patient accrual and allow for greater protocol complexity. Multicenter trials reduce the opportunity for an individual's bias to influence the conduct of the trial; they increase the likelihood for the inclusion of a more representative study population and facilitate a higher standard for data processing and analysis. Disadvantages of multicenter trials are the administrative considerations that underlie the management and administrative arrangements (eg, institutional review board, ethics committees). Considerations must be delineated for criteria for patient enrollment, diagnostic classification, and assessment of treatment outcome. These trials are inherently more costly (Box 22.2).

EVALUATION OF CLINICAL TRIALS

Many factors must be considered when evaluating a clinical trial. The most important is the clinical trial objective or aim. Whether the objectives of the trial are adequately assessed depends on the presence and extent of bias and confounding factors. Bias is a nonrandom error in a study that can alter the outcome. Types of bias to consider when evaluating a manuscript are listed in Table 22.2.

TABLE 22.2 Bias and Confounding Factors: Examination of the Literature in the Field

- Specifying and selecting the clinical trial sample
 - Popularity bias
 - Referral filter bias
 - Diagnostic or access bias
 - Wrong sample size
 - Migrator, nonrespondent, or volunteer bias
- Executing the exposure
 - Contamination, withdrawal, or compliance/therapeutic bias
 - Information bias
 - Observer bias
 - Interviewer bias
 - Use of nonvalidated instruments
 - Active control bias
- Analyzing the data
 - Post hoc significance bias
 - Fishing expeditions
- Interpreting the analysis
- Publishing the results

Specifying the sample size or number of participants in the study needed to detect a difference or an effect of a given magnitude depends on many variables. The most important is the magnitude of the effect desired or expected. Other important considerations are the desired probability of the study to identify the correct outcome (power), the variability of the variables being analyzed, the number of parts of the clinical trial, the magnitude of the placebo effect, and the number of dependent parts of the clinical trials. When determining the sample size, one must consider if the size of the treatment groups will be equal or nonequal (eg, 2:1 ratio). **An advantage of using groups of unequal size is that more information will be gained about patient responses in the larger arm. A disadvantage is the loss in study power; however, this detraction is not usually substantial if the ratios are kept under 3:1.**

Placebo Treatment Groups

Placebo treatment groups control for the psychological aspects of being in a treatment trial. They also control for adverse events being attributed to a medicine when they result from spontaneous changes in the disease or from other causes and allow a stronger interpretation of the data. Placebo treatment groups are considered ethical if the following occur:

- No standard treatment exists.
- The standard treatment has been proven ineffective.
- Standard treatment is inappropriate for the particular clinical trial.
- Placebo has been reported to be effective in treating the condition.
- The disease is mild and lack of treatment is not considered to be medically important.
- The disease process is characterized by frequent spontaneous exacerbations and remissions.

Controls Used in Clinical Trials

Control groups in clinical trials may be obtained by many different methods. Randomized control groups are the most traditional and accepted and only chance should determine who enters any of the study arms. Nonrandom control groups may also be used. Participants in these nonrandom control groups should have similar characteristics to those of the "treatment arm" and may include historical data obtained on the same patients on no therapy, the same therapy, or different therapy. Participants in the control arm may be assigned to a placebo, an active medication, or concurrent use of nontreatment; use a different dose of the same medication; or receive usual care.

Dropouts from clinical trials are inevitable. The simplest reasons may be that the participants declined to participate after enrollment or that the clinical course during the trial required a change in therapy. Whatever the reason for noncompliance or dropout, these participants should be followed because it is essential to analyze the outcomes of the groups as intent to treat. Inclusion of these data provides a conservative estimate of the differences in treatment.

Studies of Therapy

There are many different types of clinical trials; in general, they are categorized into categories or phases. Phase I trials are usually to screen the safety of the intervention or drug. These trials can be inclusive of multiple doses of a new medicine or evaluation of an old medicine in a new therapeutic area. These trials usually consist of 10 to 100 participants.

Phase II trials clarify and establish the protocol and elucidate the experimental conditions that will allow the most important phase of the trial to give a definitive result. These trials are valuable because they establish protocols and the experimental conditions that will allow the final phase of the trial to give a definitive result. They allow for the following:

- The evaluation of a variable related to the clinical pharmacology of a medicine
- Development of clinical experience by research personnel under open-label conditions before initiation of a double-blind trial

Aims of phase II trials are to assess how many people should be included in the final phase of testing, determine the endpoints of the trial, and provide preliminary estimates of effective dose and duration of treatment.

Phase III trials are for comparison to standard therapy or placebo if ethically justifiable. These trials are more commonly randomized and are regarded as the best way to obtain unbiased data.

Blinding

Blind refers to the lack of knowledge of the investigational agent by the patients, investigators, ancillary personnel, data review committees, and statisticians. Blinding is used to decrease the biases that may occur during a clinical trial. An open-label trial indicates that no blind is used. Examples of open-label trials are as follows:

- Pilot trials
- Case studies in life-threatening situations

- Unusual studies in which definitive data may be obtained (eg, coma patients)
- Clinical trials in which ethical considerations do not permit blinding

In single-blind clinical trials, either the patient or investigator is unaware of the treatment received. Single-blind trials provide a degree of control when a double-blind trial is impossible or impractical and provides a degree of assurance of the data's validity compared with open-label trials. In double-blind trials, neither the investigator nor the patient is aware of what treatment the patient is receiving. This allows for strong data interpretation if all other aspects of the trial were properly designed and conducted, the blind remained intact, the protocol was not seriously breached, the power of the trial was adequate, and the patients were compliant. Triple-blind trials are situations in which anyone who interacts with either the patient or physician is blinded. These studies allow for the strongest interpretation of data if the conditions can be met.

If blinding is to be used, then the study should be designed so that it is very difficult to break the blind. Unblinding may occur based on any of the following:

- Adverse reactions
- Lack of efficacy
- Efficacy
- Changes in laboratory values
- Errors in labeling
- Comments from unblinded study personnel
- Information presented in correspondence or reports
- Intentionally looking for clues (Box 22.3)

When to Stop a Clinical Trial

The decision to stop a clinical trial has many important ramifications. The ethical dilemmas include the needs of the next

BOX 22.3 Types of Blinds

Open label
Single blind
Double blind
Full double blind: keeping blind everyone who interacts with the patient
Full triple blind: keeping blind anyone who interacts with the patients and the investigators

eligible patient inasmuch that a participant should never be randomly assigned to an established inferior treatment. This must be balanced by the collective needs of society that terminating a trial will still result in the correct policy for the future and the need for sufficient data to change clinical practice for the better.

Early termination of a trial has its disadvantages. If the trial is stopped after recruitment of a small number of participants, the results may lack credibility. The assumed treatment difference may be the result of chance and a false-positive result. The early stopping of trials can result in imprecision and wide CIs for treatment effect. Finally, the treatment recommendation that results from stopping a trial early may be unduly enthusiastic.

Statistical stopping guidelines should be determined before the clinical trial begins. A sufficiently small *P* value for treatment difference on a trial's primary endpoint can be a guideline for when it is ethically desirable to stop a trial. It is most acceptable to have a limited number of preplanned interim analyses. Multiple repeated looks at the data can guard against the risk of a false-positive result.

For the bibliography list, log onto www.expertconsult.com (<http://www.expertconsult.com>).

BIBLIOGRAPHY

- Biggerstaff BJ, Jackson DJ: The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis, *Stat Med* 27:6093, 2008.
- Beresford SA, Weiss NS, Voigt LF, et al: Risk of endometrial cancer in relation to use of estrogen combined with cyclic progestagen therapy in Postmenopausal women, *Lancet* 349:458, 1997.
- Freidman G: *Primer of epidemiology*, ed 5, New York, 2004, McGraw-Hill.
- Giacomini MK, Cook DJ: Users' guides to the medical literature: XXIII. Qualitative research in health care A. Are the results of the study valid? Evidence-Based Medicine Working Group, *JAMA* 284:357, 2000.
- Hatch EE, Herbst AL, Hoover RN, et al: Incidence of squamous neoplasia of the cervix and vagina in women exposed prenatally to diethylstilbestrol (United States), *Cancer Causes Control* 12:837, 2001.
- Haukoos JS, Jason S: The Propensity Score, *JAMA* 314(15):1637, 2015.
- Laborde-Castérot H, Agrinier N, Thilly N: Performing both propensity score and instrumental variable analyses in observational studies often leads to discrepant results: a systematic review, *J Clin Epidemiol* 68(10):1232, 2015.
- Landis JR, Koch GG: The measurement of observer agreement for categorical data, *Biometrics* 33:159–174, 1977.
- Lasko TA, Bhagwat JG, Zou KH, et al: The use of receiver operating characteristic curves in biomedical informatics, *J Biomed Inform* 38:404, 2005.
- Li P, Stuart EA, Allison DB: Multiple imputation: a flexible tool for handling missing data, *JAMA* 314(18):1966, 2015.
- McNamee P: What difference does it make? The calculation of QALY gains from health profiles using patient and general population values, *Health Policy (New York)* 84:321, 2007.
- Montori VM, Devereaux PJ, Adhikari NK, et al: Randomized trials stopped early for benefit: a systematic review, *JAMA* 294:2203, 2005.
- Rosenbaum PR, Rubin DB: The central role of the propensity score in observational studies for causal effects, *Biometrika* 70(1):41, 1983.
- Rothman KJ, Greenland S: *Modern Epidemiology*, ed 2, Philadelphia, 1998, Lippincott-Raven.
- Spilker B: *Guide to Clinical Trials*, Philadelphia, 1991, Lippincott, Williams & Wilkins.
- Sutton A, Higgins JPT: Recent developments in meta-analysis, *Stat Med* 27:625, 2008.